

1 **Extensive positive selection drives the evolution of non-structural proteins in lineage C**

2 **Betacoronaviruses**

3

4 Diego Forni^{a*}, Rachele Cagliani^a, Alessandra Mozzi^a, Uberto Pozzoli^a, Nasser Al-Daghri^{b,c}, Mario

5 Clerici^{d,e}, Manuela Sironi^a

6

7 ^a Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy;

8 ^b Biomarker research program, Biochemistry Department, College of Science, King Saud

9 University, Riyadh, Kingdom of Saudi Arabia (KSA);

10 ^c Prince Mutaib Chair for Biomarkers of Osteoporosis, Biochemistry Department, College of

11 science, King Saud University, Riyadh, KSA;

12 ^d Department of Physiopathology and Transplantation, University of Milan, Milan, Italy;

13 ^e Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy.

14

15 Running Head: Positive Selection at MERS-CoV non-structural proteins

16

17 *Address for correspondence to Diego Forni, diego.forni@bp.lnf.it

18 Abstract word count: 249

19 Main text word count: 4998

20

21 **Abstract**

22 Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV) spread to humans via the
23 zoonotic transmission from camels. MERS-CoV belongs to the lineage C of Betacoronaviruses
24 (betaCoVs), which also includes viruses isolated from bats and hedgehogs. A large portion of the
25 betaCoV genome consists of two open reading frames (ORF1a and ORF1b) that are translated into
26 polyproteins. These are cleaved by viral proteases to generate 16 non-structural proteins (nsp1-16)
27 which compose the viral replication-transcription complex. We investigated the evolution of ORF1a
28 and ORF1b in lineage C betaCoVs. Results indicated widespread positive selection, acting mostly
29 on ORF1a. The proportion of positively selected sites in ORF1a was much higher than that
30 previously reported for the surface-exposed spike protein. Selected sites were unevenly distributed,
31 with nsp3 representing the preferential target. Several pairs of co-evolving sites were also detected,
32 possibly indicating epistatic interactions; most of these were located in nsp3. Adaptive evolution at
33 nsp3 is ongoing in MERS-CoV strains and two selected sites (G720 and R911) were detected in the
34 protease domain. Whereas position 720 is variable in camel-derived viruses, suggesting that the
35 selective event does not represent a specific adaptation to humans, the R911C substitution was only
36 observed in human-derived MERS-CoV isolates, including the viral strain responsible for the recent
37 South Korean outbreak. It will be extremely important to assess whether these changes affect host
38 range or other viral phenotypes. More generally, data herein indicate that CoV nsp3 represents a
39 major selection target and nsp3 sequencing should be envisaged in monitoring programs and field
40 surveys.

41

42

43 **Importance**

44 Both SARS-CoV and MERS-CoV originated in bats and spread to humans via an intermediate host.
45 This clearly highlights the potential for coronavirus host shifting and the relevance of understanding
46 the molecular events underling the adaptation to new host species. We investigated the evolution of
47 ORF1a and ORF1b in lineage C betaCoVs and in 87 sequenced MERS-CoV isolates. Results
48 indicated widespread positive selection, stronger in ORF1a than in ORF1b. Several selected sites
49 were found to be located in functionally relevant protein regions and some of them corresponded to
50 functional mutations in other coronaviruses. The proportion of selected sites we identified in ORF1a
51 is much higher than that for the surface-exposed spike protein. This observation suggests that
52 adaptive evolution in ORF1a might contribute to host shifts or immune evasion. Data herein also
53 indicate that genetic diversity at non-structural proteins should be taken into account when antiviral
54 compounds are developed.

55

56 Introduction

57

58 Middle East respiratory syndrome-related coronavirus (MERS-CoV,
59 http://talk.ictvonline.org/files/proposals/animal_ssna_viruses/default.aspx) was identified as the
60 causative agent of a new viral respiratory disease in Saudi Arabia in June 2012 (1). Since then, more
61 than 1,500 cases and 571 deaths have been reported worldwide (as of October 12th, 2015
62 <http://www.who.int/csr/don/12-october-2015-mers-saudi-arabia/en/>), although major outbreaks have
63 been confined to the Middle East and, more recently, to South Korea. The rate of human-to-human
64 transmission of MERS-CoV is relatively low, suggesting that a zoonotic reservoir serves as a major
65 source for transmission (2). Recent studies have indicated that MERS-CoV or a closely related virus
66 originated in bats and possibly spread to humans via transmission from dromedary camels (3).
67 Like SARS-CoV, which causes Severe Acute Respiratory Syndrome and evolved in bats as well,
68 MERS-CoV (order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae*) is a positive-
69 sense RNA (+ssRNA) virus belonging to the C lineage of the Betacoronavirus (betaCoVs) genus
70 (4). CoVs are exceptional among RNA viruses for having long (~30-kb) genomes, a feature
71 associated with a specific genome architecture and with the acquisition of an RNA 3'-to-5'
72 exoribonuclease activity (exoN) (5). About two thirds of the CoV genome consist of two large
73 overlapping open reading frames (ORF1a and ORF1b), that are translated into the polyproteins
74 pp1a and pp1ab (this latter synthesized via a -1 ribosome frameshift at the 3' end of ORF1a). These
75 polyproteins are subsequently cleaved by viral proteases to generate 16 non structural proteins
76 (nsp1 to 16), most of which compose the viral replication-transcription complex (RTC). With the
77 exception of nsp1 and 2, whose functions are poorly understood, most nsps have been characterized
78 in some detail for either MERS-CoV, SARS-CoV or mouse hepatitis virus (MHV). Thus, nsp3, a
79 large multidomain and multifunctional protein, was shown to play essential roles in the virus

80 replication cycle. In fact, the papain-like protease (PLpro) activity of nsp3 is responsible for the
81 initial processing of pp1a. Also, nsp3 together with nsp4 and 6, recruits intracellular membranes to
82 anchor the RTC and to form a reticulovesicular network of double-membrane vesicles (DVMs) and
83 convoluted membranes where viral RNA replication occurs (6). nsp5 encodes a second viral
84 protease (3C-like protease, 3CLpro) that cleaves both pp1a and pp1ab to the final nsp products.
85 nsp7-11 provide the primer-making activity and regulate the function of the main RNA-dependent
86 RNA polymerase (RdRp), this latter encoded by nsp12. Finally, nsp13-16 comprise RNA-modifying
87 enzymes, including the exoN activity in nsp14 (7, 8).

88 Viral RNA represents the major pathogen-associated molecular pattern (PAMP) recognized by the
89 host immune system during CoV infection (9). Both MERS-CoV and SARS-CoV elicit limited
90 interferon (IFN) response in most cell types, indicating that these viruses have evolved efficient
91 strategies to evade innate immune sensing and/or to block IFN induction (9). Indeed, these viruses
92 express antagonists of the IFN response including SARS-CoV ORF6, ORF3b, and nucleoprotein, as
93 well as MERS-CoV structural and accessory proteins M, ORF4a, ORF4b, and ORF5 (10-13).

94 Additional immune evasion strategies, though, rely on nsps. In fact, the enzymatic activities of
95 nsp14 and nsp16 endow the viral RNA of a 2'-O-methylated cap structure that mimics cellular
96 mRNAs and avoids activation of the innate immunity receptors RIG-I and MDA5 (9). In analogy to
97 the exoN and endoribonucleases expressed by other viruses such as Lassa Fever Virus and
98 Pestivirus (14, 15), the ribonuclease activities of nsp14 and nsp15 are also thought to play a role in
99 immune escape by digesting RNA PAMPs (9). Moreover, suppression of IFN responses is mediated
100 by the PLpro in nsp3 through its deubiquitinating and deISGylating activities (16, 17), as well as by
101 nsp1. The latter inhibits IFN-dependent signaling by decreasing the phosphorylation levels of
102 STAT1 and suppresses host protein synthesis (18, 19). Finally, PLpro was shown to physically
103 interact with TRAF3, TBK1, IKK ϵ , STING, and IRF3, which represent key cellular components for

104 IFN response (20).

105 Therefore, the information encoded by CoV ORF1a and ORF1b is essential for viral replication and
106 for immune evasion. For these reasons, inhibitors that interfere with viral enzymatic activities (e.g.
107 proteases, RdRp) are regarded as promising candidates for therapeutic intervention (21).
108 From an evolutionary standpoint, different observations suggest that nsps may represent targets of
109 natural selection. First, genes encoding molecules that directly interact with the host immune
110 system are thought to be preferential targets of natural selection as a consequence of host-pathogen
111 arms races (22). Second, adaptation to new hosts in other RNA viruses has been associated to
112 selective changes in polymerase genes (23, 24). Finally, the acquisition of a complex replication
113 machinery is evolutionary linked to genome expansion in *Nidovirales* (5). Nonetheless,
114 evolutionary studies have mainly focused on the analysis of betaCoV spike proteins, as these are
115 surface exposed and represent major determinants of host range and tissue tropism (25). Herein we
116 investigated the evolution of ORF1a and ORF1b in MERS-CoV and in lineage C betaCoVs isolated
117 from bats and hedgehogs. Results indicate widespread positive selection, stronger in ORF1a; within
118 this region, nsp3 represents a preferential selection target and adaptive evolution is ongoing in
119 MERS-CoV strains circulating in the current outbreak.

120 **Material and Methods**

121

122 **Sequences and alignments**

123 ORF1a/ORF1b sequences for 7 lineage C betaCoVs and 87 MERS-CoV strains (available as of July
124 2015) were retrieved from the NCBI database; a list of accession numbers for the complete
125 genomes is provided in Table S1.
126 Alignment errors are common when divergent sequences are analyzed and can affect evolutionary
127 inference. Thus, we used PRANK (26) to generate multiple sequence alignments and GUIDANCE

(27) for filtering unreliably aligned codons (i.e. we masked codons with a score <0.90), as suggested (28). The alignments were screened for the presence of recombination events using two methods based on distinct data features: 1) GARD (Genetic Algorithm Recombination Detection) (29) uses phylogenetic incongruence among segments in the alignment to detect the best-fit number and location of recombination breakpoints; the statistical significance of putative breakpoints is then evaluated through Kishino-Hasegawa (HK) tests; 2) GENECONV (30) tests for significant clustering of substitutions along sequences; statistical significance is assessed through permutation with multiple-comparison correction. For both methods recombination breakpoints were considered significant if the p value was < 0.05 . No breakpoint was detected in any analysis.

138

139 **Detection of positive selection**

140

Gene trees were generated by maximum-likelihood using the program phyML with a GTR plus gamma-distributed rates model and 4 substitution rate categories(31). Positive selection can be defined when the nonsynonymous/synonymous rate ratio (ω) is higher than 1; to analyze the presence of episodic positive selection in lineage C betaCoVs viruses we applied the branch-site test (32) from the PAML suite (33). The test is based on the comparison between two nested models: a model (MA) that allows positive selection on one or more lineages (called foreground lineages) and a model (MA1) that does not allow such positive selection. Twice the difference of likelihood for the two models ($\Delta\ln$) is then compared to a χ^2 distribution with one degree of freedom (32). A false discovery rate correction was applied to take into account a multiple hypothesis issue generated by analyzing different branches on the same phylogeny (34). When the likelihood ratio test suggested the action of positive selection, the Bayes Empirical Bayes (BEB)

152 analysis was used to evaluate the posterior probability that each codon belongs to the site class of
153 positive selection on the foreground branch.

154 BUSTED (branch-site unrestricted statistical test for episodic diversification) (35) is a recently
155 developed software designed to describe episodic positive selection that is acting on specific
156 branches in the phylogeny at a proportion of sites within the alignment. An alternative model that
157 allows the action on positive selection on foreground branches is compared with a null model that
158 doesn't allow $\omega > 1$. Twice the $\Delta \ln$ of the two models is then compared to a χ^2 distribution (degrees
159 of freedom=2); if the null model is rejected, at least one site is under positive selection on the
160 foreground branch(es). To detect selection at individual sites, twice the difference of the likelihood
161 for the alternative and the null model at each site is compared to a χ^2 distribution (degree of
162 freedom=1). BUSTED is implemented in the HYPHY package (36).

163 Conservatively, we considered a site as selected if it showed a p value ≤ 0.05 in BUSTED and a
164 posterior probability ≥ 0.90 in the BEB analysis.

165 The site models implemented in PAML were applied for the analysis of nsp3 sequences from
166 MERS-CoVs isolates. To detect selection, two different pairs of nested site models (M1a/M2a and
167 M7/M8) were fitted to the data (33); the M2a and M8 allow a class of sites to evolve with $\omega > 1$,
168 whereas M1a and M7 do not. Positively selected sites were identified using the BEB analysis (from
169 model M8) (37). Sites were validated using MEME (38) (with a cutoff ≤ 0.1), which allows the
170 distribution of ω to vary from site to site and from branch to branch at a site.

171 MEME (38) analyses were performed through the DataMonkey server (39).

172

173

174 **Detection of co-evolving sites**

175 To detect co-evolving sites in the ORF1a and ORF1b alignments we applied two different methods:

176 BGM (Bayesian Graphical Model)-Spidermonkey (40) and the Mutual Information Server To Infer
177 Coevolution (MISTIC) (41). Spidermonkey is a tool implemented in the HYPHY package that
178 identifies co-evolving sites from an alignment of coding sequences; a BGM is used to evaluate the
179 connection among codons (represented by the nodes of the network). Significant statistical
180 associations between nodes are indicated by the edges of the network, suggesting functional or
181 structural interactions between codons.

182 MISTIC estimates the relationship between two or more position in an alignment. The co-
183 evolutionary association is estimated by Mutual Information (MI), that evaluates how much the
184 information from the aminoacid at the first position can help to predict the aminoacid identity at the
185 second position.

186 For BGM-Spidermonkey sites were filtered based on a minimum count of 4 substitutions across the
187 phylogeny. To be conservative, we considered a pair of residues as co-evolving if they showed a
188 posterior probability >0.75. This threshold corresponds to 0.02% and 1.42% of all analyzed site
189 pairs in ORF1a and ORF1b, respectively. Likewise, for MISTIC site pairs were required to display
190 a MI rank higher than the 99th percentile calculated using all MI scores from the alignment. Pairs of
191 sites exceeding the thresholds for both methods were declared to be co-evolving.

192

193 **Membrane topology, glycosylation site predictions, and 3D structure mapping**

194 The membrane protein topology for MERS-CoV nsp3 and nsp4 was predicted by using TMHMM
195 (<http://www.cbs.dtu.dk/services/TMHMM/>) (42). N-Glycosylation sites were predicted with
196 NetNGlyc (<http://www.cbs.dtu.dk/services/NetNGlyc/>), a program that uses artificial neural
197 networks to examine the sequence context of Asn-X-Ser/Thr motifs.

198 Sites were mapped onto structures using PyMOL (The PyMOL Molecular Graphics System,
199 Version 1.5.0.2 Schrödinger, LLC).

200 Results

201

202 nsp3 in ORF1a is a major selection target in betaCoVs

203

204 The lineage C of betaCoVs includes two bat species closely related to MERS-CoV, namely Ty-
205 BatCoV HKU4 and Pi-BatCoV HKU5, isolated from the lesser bamboo bats (*Tylonycteris*
206 *pachypus*) and Japanese pipistrelles (*Pipistrellus abramus*), respectively (4). Additional viruses
207 belonging to the lineage C of betaCoVs have been described in bats (BtCoV/133, BtVs-
208 BetaCoV/SC2013) and hedgehogs (Hedgehog coronavirus, EriCoV) (21, 43-45). Recently, a virus
209 belonging to the same species as MERS-CoV was isolated in *Neoromicia* bats (NeoCoV) (46). To
210 investigate the evolutionary history of ORF1a, we obtained sequence information for these viruses
211 and for 6 MERS-CoV strains isolated from either humans or camels and belonging to the major
212 groups described to date (47) (Fig. 1). The sequence alignment was pruned of unreliably aligned
213 codons (see Material and Methods), a procedure that resulted in the masking of the almost entire
214 acidic domain in nsp3. Indeed, this region was previously shown to be highly divergent among
215 CoVs (48). We next analyzed the alignment for the presence of recombination breakpoints using
216 GARD (Genetic Algorithm Recombination Detection) (29) and GENECONV (30). No evidence of
217 recombination was detected.

218 The phylogenetic tree of ORF1a obtained with phyML was consistent with previously reported ones
219 (46, 47). An estimate of the extent of functional constraint along ORF1a was obtained by
220 identification of negatively selected sites (total number =903) and calculation of their distribution
221 among nsps. This analysis indicated that the average fraction of negatively selected sites is ~0.24,
222 with weakest selection in nsp1 and strongest constraint in nsps6-9 (Fig. 1).

223 Evidence of episodic positive selection along the internal branches of the ORF1a phylogeny was

224 searched for using branch-site tests. Specifically, we applied two different methods: the branch-site
225 unrestricted statistical test for episodic diversification (BUSTED) (35) and the maximum-likelihood
226 models (MA/MA1) implemented in the PAML suite (33). These two approaches rely on different
227 assumptions of ω (nonsynonymous/synonymous rate ratio) variation among branches. To be
228 conservative, episodic positive selection at each tested branch was declared only if statistically
229 significant support was obtained with both methods. Using this criterion, we found 4 branches with
230 evidence of episodic selection (Fig. 1, Table 1). Selected sites along these branches were identified
231 using the Bayes empirical Bayes (BEB) procedure from model MA and with BUSTED; again, only
232 sites detected by both methods were considered. A total of 55 selected sites were detected; these
233 were scattered along the entire ORF1a and located in different nsps, with the exclusion of nsp7 and
234 nsp9, which were not targeted by selection (Fig. 1, Table S2). To determine whether any nsp
235 represented a preferential target of episodic positive selection, we performed random uniform
236 sampling (i.e. assuming a random distribution of selected sites, we calculated the likelihood of
237 identifying a number of sites equal to or higher than the one we observed for each nsp). This
238 analysis indicated that nsp3 was preferentially targeted by episodic selection (Bonferroni corrected
239 p value= 0.010) during the evolution of MERS-related CoVs.

240 Finally, we searched for evidence of co-evolution between sites in the ORF1a alignment. To this
241 aim, we applied BGM-Spidermonkey (40) and MISTIC (41) (see Material and Methods for details).
242 Six pairs of co-evolving sites were detected by both methods. Most site pairs were located in nsp3
243 (Fig. 1, Table S3).

244

245 **Weaker selective pressure for ORF1b**

246

247 The same approach described above was applied to ORF1b sequences. The degree of constraint in

248 nsp12-16 was comparable to that observed for ORF1a (Fig. 1A). Statistical support of episodic
249 positive selection was detected for 2 branches only (Fig. 1B, Table 1) and fewer selected sites were
250 detected compared to ORF1a. Most sites were targeted by selection on the branch leading to bat
251 CoVs (HKU5, BtCoV/133, HKU4) (Fig. 1, Table S2). No nsp resulted a preferential selection
252 target. One pair of co-evolving sites was detected by MISTIC and BGM-Spidermonkey (Fig. 1A,
253 Table S3).

254

255 **Ongoing selection at nps3 in MERS-CoV**

256

257 Given the results above, we investigated whether positive selection also occurred at nsp3 during the
258 recent evolution of MERS-CoV. To this purpose, we retrieved 87 sequences from MERS-CoVs
259 isolated from camels or humans (available strains as of July 2015) (Table S1). No recombination
260 breakpoint was detected by either GARD or GENECONV and the *codeml* site models were applied
261 (33). Results showed that models of gene evolution that allow a class of codons to evolve with $\omega > 1$
262 (NSsite models M2a and M8) better fit the data than the neutral models (NSsite models M1a and
263 M7), strongly supporting the action of positive selection (Table 2). Positively selected sites in nsp3
264 were identified using two methods, the BEB procedure implemented in M8 and MEME (Mixed
265 Effects Model of Evolution). Two sites (G720 and R911) were detected by both methods (Table 2,
266 Fig. 2). Four different residues are observed at position 720 in MERS-CoV sequences, and three of
267 them are present in viruses isolated from camels (Fig. 2), suggesting that adaptive evolution at this
268 site started before the spread to humans. Conversely, the R911C substitution only occurs in viruses
269 derived from humans: all of them belong to group 3 and include the virus responsible for the recent
270 South Korean outbreak (MERS-CoV/KOR/KNIH) (Fig. 2).

271

272 **Positively selected sites in the context of betaCoV biology**

273

274 To shed light into the functional role of selected sites we exploited available structural, genetic, and
275 biochemical data for nsps (mainly obtained for SARS-CoV and MHV), as well as *in silico*
276 prediction of transmembrane helices and glycosylation sites.

277 Most positively selected sites were located in nsp3 (Fig. 3A). In the PLpro domain, four of the
278 positively selected sites, including two that are selected in the viral strains from the current MERS-
279 CoV epidemic, were clustered in a spatially confined region opposed to the catalytic crevice. This
280 region corresponds to the “palm” of the right-handed architecture of the protease, whereas three
281 additional sites are located on the “fingers” (Fig. 3A). One of these three surface exposed sites,
282 Q830, is part of a non-canonical nsp5 cleavage site, unique for MERS-CoV, that could contribute to
283 the formation of new cleavage products (49) (Fig. 3A).

284 Additional sites were identified in the other domains of the protein (Fig. 3A). One positively
285 selected site in the Ubl1 domain (C91) precedes a conserved acidic loop that when mutated in
286 SARS-CoV determines a lethal phenotype (50). Among sites in the ADPR domain, none was
287 located within the ADP-ribose binding pocket (Fig. 3A) (51). As for the nucleic-acid binding
288 domain (NAB) (Fig. 3A), one of the positively selected site (T1080) is part of the α 2 helix, that
289 together with the loop preceding the β 6 strand, forms a structure possibly interacting with ssRNA
290 via a charged patch (52).

291 Finally, positively selected sites were found in the so-called transmembrane region of nsp3 (Fig. 3).

292 We performed an *in silico* prediction of transmembrane helices and glycosylation sites. As
293 previously observed for MHV and SARS-CoV (53), the topology model did not fit the general
294 Nendo/Cendo shared structure and was inconsistent with the location of the glycosylation site.
295 Thus, in analogy to other CoVs, some hydrophobic regions predicted to be transmembrane are

296 unlikely to span the lipid bilayer. Intriguingly, we identified two positively selected sites in the
297 luminal loop; one of them (A1386) is located between the first two conserved cysteine residues
298 (Fig. 3B) of the zinc finger motif (ZF). Other positively selected sites were found in the third
299 hydrophobic region (that may or may not span the membrane).

300 In nsp4, four transmembrane regions were predicted (in SARS-CoV and MHV all of these are
301 membrane-spanning) (54) (Fig. 3C). One of the positively selected sites is located on the large
302 luminal loop 1 (Fig. 3C) and does not affect (or introduces) a predicted glycosylation site. Variants
303 in the same loop of MHV nps4 were shown to alter DVM morphology or number, irrespective of
304 their effect on glycosylation (55). Two additional positively selected sites (G453 and A479) are
305 located in the C-terminal domain of the protein (Fig. 3C). This domain is cytosolic and interacts
306 with viral and host proteins (56).

307 We also detected 4 positively selected sites in the other viral protease, 3CLpro (nsp5) (Fig. 4).
308 Interestingly, two of them (H8 and V132) were proposed to affect the correct domain orientation
309 required for dimerization of MERS-CoV nsp5 (57). Moreover, site-directed mutagenesis of 3CLpro
310 from MHV indicated that the two corresponding sites affect protease activity in a temperature-
311 dependent fashion (58, 59) (Fig. 4).

312 Concerning selected sites in nsp10 and ORF1b, interesting findings relate to the nps10-nsp14 and
313 nsp10-nsp16 interactions. Based on the crystal structure of the corresponding SARS-CoV protein
314 complexes, one of the positively selected sites in nsp16 (K249) is located at the direct interaction
315 surface with nsp10 (Fig. 5A). Also, the only positively selected site we detected in nsp10 (S61) is
316 involved in the interaction with both nsp16 (60) and nsp14 (61) (Fig. 5). In fact, mutation of residue
317 61 in SARS-CoV nsp10 (A61V mutant) strongly reduces interaction with nsp14. In MHV and
318 SARS-CoV, S61 is located on an exposed loop that also includes Q65 (Fig. 5); mutation of Q65
319 results in a temperature sensitive (ts) phenotype in MHV and, through a poorly understood

320 mechanism, disrupts nsp5 function (62). This observation suggests that this loop may be involved in
321 interaction with other nsps, or that aminoacid changes in this loop can modify nsp10 conformation
322 (62).

323 Additional selected sites in nsp14 and 16 do not involve the contact interface with nsp10. In nsp14,
324 T228 is located within one of the zinc fingers of the protein (Fig. 5); these domains are essential for
325 the ExoN proofreading activity of nsp14 (63). Another selected residue (S284) flanks a position
326 (F286 in SARS, highly conserved in coronaviruses) that is involved in the hydrophobic interaction
327 between the ExoN and N7-methyltransferase domain (63).

328 As for nsp16, the T151 selected site is in close proximity to a residue (L153) (Fig. 5) that originates
329 a ts phenotype in MHV; specifically, the mutation has an effect on RNA synthesis (64). Finally, the
330 positively selected site 138 (Asn in MERS-CoV) is located on a solvent-exposed flexible loop (not
331 present in the structure) on the RNA binding groove; this loop may be involved in interaction with
332 RNA (65).

333

334 Discussion

335

336 In recent years, the emergence of SARS-CoV and MERS-CoV as dangerous zoonoses stirred great
337 interest in the ecology and evolution of coronaviruses. Both viruses originated in bats and spread to
338 humans via an intermediate host (3, 66). This clearly highlights the potential for CoV host shifting
339 and the relevance of understanding the molecular events underling the adaptation to new species.

340 Since the identification of MERS-CoV, a number of related viruses were isolated from bats (and
341 other mammals) all over the world, suggesting a wide distribution of betaCoVs. Unfortunately, most
342 of these studies reported partial viral sequences, hampering a fully resolved phylogenetic analysis
343 (67-69). Herein we only included viral species/strains with complete information for ORF1a and

344 ORF1b, and the sequences we analyzed are relatively divergent. This may introduce a high false
345 positive rate in the inference of positive selection due to two major issues: the unreliability of
346 sequence alignments and the saturation of substitution rates. To circumvent these possible problems
347 we adopted a stringent alignment filtering criterion and we computed substitution rates over
348 phylogenies. This latter procedure allows breaking of long branches, resulting in improved rate
349 estimation. In fact, branch-site tests are relatively insensitive to the saturation issue (70). Moreover,
350 we supported all claims of positive selection by the combined use of two methods. Although this
351 may have resulted in a loss of power (that is intrinsically low for branch site tests (32)), we were
352 able to detect several positively selected sites along different branches of the phylogenies. In both
353 ORF1a and ORF1b episodic selection was particularly strong on the branch leading to bat CoVs
354 (HKU4, HKU5, and BtCoV133), possibly reflecting specific characteristics of these host species
355 (large population sizes, high seroprevalence, and wide geographic distributions).

356 Recently, a study of the spike (S) protein in MERS-CoV-related viruses analyzed a very similar set
357 of viral sequences as the ones herein and detected positive selection at 9 sites (25). The proportion
358 of selected sites we identified in the ORF1a region is much higher than that for the S gene. Albeit
359 counter-intuitive, as the S protein is exposed on the viral surface and functions as a central
360 determinant of host range and of antibody response, this observation indicates that ORF1a
361 represented a major target of positive selection during the evolution of lineage C betaCoVs.

362 Therefore, adaptive evolution in this region might contribute to host shifts or immune evasion. In
363 line with these results, studies on avian influenza A viruses showed that viral surface proteins are
364 not the sole determinants of host range. Conversely, adaptation of avian influenza to mammalian
365 hosts was found to be critically dependent on changes in the polymerase genes (23, 24). Recently,
366 an extended analysis indicated that the adaptation of avian influenza virus to swine was
367 accompanied by substitutions in almost all viral genes. Mutation accrual continued for a long time

368 after the host shift, suggesting that multiple mutations progressively optimize viral fitness in the
369 new host (71); many of these mutations were suggested to represent compensatory or epistatic
370 changes (71). Interestingly, we also found evidence of co-evolution between site pairs, possibly
371 suggesting epistatic interactions. The vast majority of co-evolving sites were located in nsp3, in line
372 with the view that epistatically interacting substitutions are enriched in protein regions undergoing
373 adaptive evolution (72). Epistasis is very common in viruses and is thought to play an important
374 role in the evolution of immune evasion, host shifts, and drug resistance (73, 74); this latter is likely
375 not at play in the case of MERS-CoV and related viruses. Previous studies on RNA viruses have
376 mainly focused on the effect of epistatic interaction on the emergence of drug resistance (e.g.
377 influenza virus hemagglutinin and neuraminidase, HCV NS3 protease, HIV-1 protease and reverse
378 transcriptase) or antibody escape (influenza virus hemagglutinin and neuraminidase) (75-78). Little
379 information is available on the possible role of co-evolving sites in the context of innate immunity
380 or adaptation to new hosts. Nonetheless, epistatic interaction between two sites in the Chikungunya
381 virus E1 envelope glycoprotein underlies the ability of the pathogen to infect a new mosquito vector
382 (79). These observations suggest that intra- and inter-gene epistasis contributes to determine the
383 evolutionary trajectories of viral species whenever the environment (broadly defined, including the
384 host species) changes (74). The co-evolving sites we identified are located in distinct domains of
385 nsp3 and the pair in nsp12 and nsp14 does not involve residues at the interaction surface between
386 the two proteins. Thus, we are presently unable to infer the molecular mechanism underlying their
387 interaction. The generation of mutant viruses or recombinant viral proteins carrying different amino
388 acids at co-evolving site pairs will be necessary to address the nature of their interaction and the
389 effect on the viral phenotype. We note, however, that long-range interaction are likely to be
390 common in ORF1a and ORF1b, both at the intra- and inter-protein level, as demonstrated by
391 experiments with SARS-CoV and MHV mutants (58, 62). For instance, a temperature sensitive (ts)

392 mutation in MHV nsp10 affects viral replication by blocking the activity of nsp5 (62). Also, ts
393 mutations in MHV nsp5 can be rescued by second-site suppressor changes located in physically
394 distant protein regions (58).

395 The CoV RTC is extremely complex and comprises a number of enzymatic activities that constitute
396 a unique repertoire among RNA viruses. Several molecular details on the assembly and functioning
397 of the RTC are still missing and the mechanisms underlying the phenotypic effects of specific
398 changes are often poorly understood. For instance, Stokes and coworkers described a conservative
399 mutation (M575I) located in the N-terminal portion of the ADPR domain of MHV nsp3 that
400 determines a ts phenotype with impaired RNA synthesis (80). How the mutation exerts its effects,
401 though, remains unknown. We thus interpreted the effect of positively selected sites in light of what
402 is known of CoV biology, and we compared the location of selected sites to the few substitutions (in
403 either SARS-CoV or MHV) with a known effect, as detailed above. Clearly, the selected sites we
404 identified herein lend themselves to experimental testing to assess their impact on BetaCoV
405 phenotypes. Indeed, evolutionary analyses can provide information on the location and nature of
406 adaptive changes, thus highlighting the presence of functional genetic variants. For instance, the
407 major selection target, nsp3, has multiple enzymatic functions. It would be extremely interesting to
408 use site-directed mutagenesis and biochemical analyses to assess whether the selected sites in the
409 PLpro domain affect the enzyme's specificity not only towards the viral polyprotein, but also in
410 terms of deubiquitination and deISGylation activities, which, in turn, may modulate the viral ability
411 to evade host immune responses (17). Likewise, expression of mutant nsp3 and nsp4 proteins
412 carrying different aminoacids at the positively selected sites in the large luminal loops will be
413 instrumental to determine whether, as shown for other CoVs (54, 55), these changes affect the
414 formation of membrane rearrangements onto which the RTC assembles. Ultimately, the generation
415 of mutant CoVs followed by in vitro infection will clarify the effect of specific changes on viral

416 fitness, at least in cell cultures.

417 It is noteworthy that we found nsp3 to represent a preferential target of positive selection and that

418 the adaptive process is ongoing in circulating MERS-CoV strains. A previous genome-wide study of

419 positive selection in MERS-CoV detected only one positively selected site in the *S* gene (2);

420 however, fewer strains than those we analyzed herein were available at the time of the study and the

421 authors did not include sequences obtained from camels (2). The two selected sites we identified are

422 located in the PLpro domain of nsp3 and, together with two sites selected in bats, map to the “palm”

423 portion of the right-handed structure of the protease. The palm also accommodates the catalytic

424 triad, but the selected sites are located at the opposite side of the crevice, suggesting that they may

425 exert an effect by altering the conformational structure of the protease or via interaction with other

426 viral components or host proteins. In addition to its role in viral replication PLpro functions as a

427 multitasking inhibitor of IFN responses and physically associates with several host innate immunity

428 molecules (20). Because antagonism of the host immune system is considered a major driver of

429 evolutionary change in viruses (81), the G720K variant is an excellent candidate as a modulator of

430 host responses. However, variability at position 720 is also observed among viruses isolated from

431 camels, suggesting that the selective event ensued in these animals or in a previous host and does

432 not represent a specific adaptation to humans. A similar observation was previously reported for the

433 positively selected site in the *S* protein (25). Conversely, variation at the second selected site (R911)

434 in PLpro was only observed for viruses isolated from human patients. Although this finding may

435 simply reflect the sparse sampling of camel-derived viruses, adaptation of a zoonotic virus to a new

436 host is expected to result in selective events that optimize viral fitness in terms of replication

437 efficiency, transmissibility, and immune evasion (73).

438 In general, deeper understanding of the adaptive events that underlie host shifts in CoVs and other

439 viruses will be pivotal to predict and prevent future zoonoses. Bats alone host a variety of CoVs that

440 represent potential threats to human health (67-69). Data herein suggest that monitoring programs
441 and field surveys of CoV diversity and prevalence should envisage molecular characterization of
442 nsp3.

443

444

445 **Funding information**

446 NAD is supported by the Deanship of Scientific Research, Prolific Research Group Program (PRG-
447 1436-15), Vice Rectorate for Graduate Studies and Scientific Research in King Saud University
448 (KSU), Riyadh, Saudi Arabia. The funder had no role in study design, data collection and
449 interpretation, or the decision to submit the work of publication.

450

451

452

453

454

455

456

457

458

459

460

461

462 **Figure Legends**

463 **Figure 1. (A)** Schematic representation of ORF1a/ORF1b and their nsps products. nsps are colored
464 in hues of blue depending on the percentage of negatively selected sites. nsp11 is shown in gray
465 because it was not analyzed (because it is too short). Positively selected sites are represented by
466 triangles, with color corresponding to the selected branch in the phylogeny (see panel B). Co-
467 evolving sites are shown below the nsp structure, with different symbols indicating each pair of co-
468 evolving sites. **(B)** Maximum likelihood phylogenies for ORF1a (left) and ORF1b (right) in lineage
469 C betaCoVs. Branches set as foreground lineages in independent branch-site tests are highlighted
470 with different colors and numbered. Thick branches yielded statistically significant evidence of
471 positive selection. Branch length is proportional to synonymous substitution rate (dS). Co-evolving
472 sites are also reported, with different symbols as indicated in panel A. Positions are relative to each
473 nsp, see also Table S3.

474

475 **Figure 2.** Maximum likelihood phylogeny for nsp3 sequences in a subset of isolates representing
476 MERS-CoV major groups. The amino acid alignment of the region surrounding the two positively
477 selected sites (magenta) in MERS-CoV isolates is also shown. Asterisks denote viruses isolated
478 from dromedary camels.

479

480 **Figure 3. (A)** Representation of nsp3 domain architecture. Positively selected sites are indicated by
481 triangles, co-evolving sites with symbols (see Figure 1 legend). In the enlargements, positively
482 selected sites were mapped onto known domain 3D-structures of MERS-CoV or SARS-CoV (PDB
483 IDs: 2GRI, 3EWR, 4RNA, and 2K87). The acidic domain is shown in gray because it was not
484 analyzed (see text). Topology maps and probability diagrams of transmembrane helices for MERS-
485 CoV nsp3 transmembrane domain **(B)** and MERS-CoV nsp4 **(C)**. The conserved cysteine residues

486 and the predicted N-glycosylation sites are mapped onto the luminal loops. Color codes are as in
487 Figure 1; yellow denotes protein regions or sites known to be functional and mentioned in the text.
488

489 **Figure 4. (A)** Structure of the dimeric form of MERS-CoV nsp5 (PDB ID: 4YLU). Positively
490 selected sites affecting the dimerization process are labeled onto the structure. Color codes are as in
491 Figure 1. Catalytic residues are in yellow. **(B)** Amino acid alignment of the nsp5 regions
492 surrounding selected sites. Residues that confer a temperature sensitive phenotype when mutated in
493 MHV are underlined (see text).

494

495 **Figure 5.** Ribbon representation of SARS-CoV nsp10-nsp16 complex (PDB ID: 2XYQ) **(A)**, and
496 SARS-CoV nsp10-nsp14 complex (PDB ID: 5C8U) **(B)**. Positively selected sites in lineage C
497 betaCoVs are shown in green (see Figure 1), residue involved in inter- or intra-protein interactions
498 are shown in yellow. An amino acid alignment of the nsp10 exposed loop containing S61 and Q65
499 is also shown. Functional residues in MHV/SARS-CoV are underlined (see text).

500

501

502

503

504

505

506

507

508

509

510 **References**

- 511 1. **Zaki, A. M., S. van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier.** 2012.
512 Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.*
513 **367**:1814-1820. doi: 10.1056/NEJMoa1211721.
- 514 2. **Cotten, M., S. J. Watson, A. I. Zumla, H. Q. Makhdoom, A. L. Palser, S. H. Ong, A. A. Al**
515 **Rabecah, R. F. Alhakeem, A. Assiri, J. A. Al-Tawfiq, A. Albarrak, M. Barry, A. Shibl, F. A.**
516 **Alrabiah, S. Hajjar, H. H. Balkhy, H. Flemban, A. Rambaut, P. Kellam, and Z. A. Memish.**
517 2014. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus.
518 *Mbio.* **5**:10.1128/mBio.01062-13. doi: 10.1128/mBio.01062-13.
- 519 3. **Al-Tawfiq, J. A., and Z. A. Memish.** 2014. Middle East respiratory syndrome coronavirus:
520 transmission and phylogenetic evolution. *Trends Microbiol.* **22**:573-579. doi: S0966-
521 842X(14)00157-7.
- 522 4. **de Groot, R. J., S. C. Baker, R. S. Baric, C. S. Brown, C. Drosten, L. Enjuanes, R. A.**
523 **Fouchier, M. Galiano, A. E. Gorbalenya, Z. A. Memish, S. Perlman, L. L. Poon, E. J. Snijder,**
524 **G. M. Stephens, P. C. Woo, A. M. Zaki, M. Zambon, and J. Ziebuhr.** 2013. Middle East
525 respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *J.*
526 *Virol.* **87**:7790-7792. doi: 10.1128/JVI.01244-13.
- 527 5. **Lauber, C., J. J. Goeman, C. Parquet Mdel, P. T. Nga, E. J. Snijder, K. Morita, and A. E.**
528 **Gorbalenya.** 2013. The footprint of genome architecture in the largest genome expansion in RNA
529 viruses. *PLoS Pathog.* **9**:e1003500. doi: 10.1371/journal.ppat.1003500.
- 530 6. **Hagemeijer, M. C., P. J. Rottier, and C. A. de Haan.** 2012. Biogenesis and dynamics of the
531 coronavirus replicative structures. *Viruses.* **4**:3245-3269. doi: 10.3390/v4113245.
- 532 7. **Neuman, B. W., P. Chamberlain, F. Bowden, and J. Joseph.** 2014. Atlas of coronavirus
533 replicase structure. *Virus Res.* **194**:49-66. doi: 10.1016/j.virusres.2013.12.004.

- 534 8. **Hagemeijer, M. C., P. J. Rottier, and C. A. de Haan.** 2012. Biogenesis and dynamics of the
535 coronavirus replicative structures. *Viruses*. **4**:3245-3269. doi: 10.3390/v4113245.
- 536 9. **Kindler, E., and V. Thiel.** 2014. To sense or not to sense viral RNA--essentials of coronavirus
537 innate immune evasion. *Curr. Opin. Microbiol.* **20**:69-75. doi: 10.1016/j.mib.2014.05.005.
- 538 10. **Kopecky-Bromberg, S. A., L. Martinez-Sobrido, M. Frieman, R. A. Baric, and P. Palese.**
539 2007. Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and
540 nucleocapsid proteins function as interferon antagonists. *J. Virol.* **81**:548-557. doi: JVI.01782-06.
- 541 11. **Frieman, M., B. Yount, M. Heise, S. A. Kopecky-Bromberg, P. Palese, and R. S. Baric.**
542 2007. Severe acute respiratory syndrome coronavirus ORF6 antagonizes STAT1 function by
543 sequestering nuclear import factors on the rough endoplasmic reticulum/Golgi membrane. *J. Virol.*
544 **81**:9812-9824. doi: JVI.01012-07.
- 545 12. **Yang, Y., L. Zhang, H. Geng, Y. Deng, B. Huang, Y. Guo, Z. Zhao, and W. Tan.** 2013. The
546 structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of Middle East respiratory
547 syndrome coronavirus (MERS-CoV) are potent interferon antagonists. *Protein Cell.* **4**:951-961. doi:
548 10.1007/s13238-013-3096-8.
- 549 13. **Yang, Y., F. Ye, N. Zhu, W. Wang, Y. Deng, Z. Zhao, and W. Tan.** 2015. Middle East
550 respiratory syndrome coronavirus ORF4b protein inhibits type I interferon production through both
551 cytoplasmic and nuclear targets. *Sci. Rep.* **5**:17554. doi: 10.1038/srep17554.
- 552 14. **Qi, X., S. Lan, W. Wang, L. M. Schelde, H. Dong, G. D. Wallat, H. Ly, Y. Liang, and C.**
553 **Dong.** 2010. Cap binding and immune evasion revealed by Lassa nucleoprotein structure. *Nature*.
554 **468**:779-783. doi: 10.1038/nature09605.
- 555 15. **Python, S., M. Gerber, R. Suter, N. Ruggli, and A. Summerfield.** 2013. Efficient sensing of
556 infected cells in absence of virus particles by plasmacytoid dendritic cells is blocked by the viral
557 ribonuclease E(rns.). *PLoS Pathog.* **9**:e1003412. doi: 10.1371/journal.ppat.1003412.

- 558 16. **Baez-Santos, Y. M., S. E. St John, and A. D. Mesecar.** 2015. The SARS-coronavirus papain-
559 like protease: structure, function and inhibition by designed antiviral compounds. *Antiviral Res.*
560 **115**:21-38. doi: 10.1016/j.antiviral.2014.12.015.
- 561 17. **Baez-Santos, Y. M., A. M. Mielech, X. Deng, S. Baker, and A. D. Mesecar.** 2014. Catalytic
562 function and substrate specificity of the papain-like protease domain of nsp3 from the Middle East
563 respiratory syndrome coronavirus. *J. Virol.* **88**:12511-12527. doi: 10.1128/JVI.01294-14.
- 564 18. **Kamitani, W., C. Huang, K. Narayanan, K. G. Lokugamage, and S. Makino.** 2009. A two-
565 pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nat. Struct.*
566 *Mol. Biol.* **16**:1134-1140. doi: 10.1038/nsmb.1680.
- 567 19. **Wathelet, M. G., M. Orr, M. B. Frieman, and R. S. Baric.** 2007. Severe acute respiratory
568 syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an attenuated
569 strain. *J. Virol.* **81**:11620-11633. doi: JVI.00702-07.
- 570 20. **Chen, X., X. Yang, Y. Zheng, Y. Yang, Y. Xing, and Z. Chen.** 2014. SARS coronavirus
571 papain-like protease inhibits the type I interferon signaling pathway through interaction with the
572 STING-TRAF3-TBK1 complex. *Protein Cell.* **5**:369-381. doi: 10.1007/s13238-014-0026-3.
- 573 21. **Chan, J. F., S. K. Lau, K. K. To, V. C. Cheng, P. C. Woo, and K. Y. Yuen.** 2015. Middle East
574 respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease.
575 *Clin. Microbiol. Rev.* **28**:465-522. doi: 10.1128/CMR.00102-14.
- 576 22. **Sironi, M., R. Cagliani, D. Forni, and M. Clerici.** 2015. Evolutionary insights into host-
577 pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**:224-236. doi:
578 10.1038/nrg3905.
- 579 23. **Hatta, M., P. Gao, P. Halfmann, and Y. Kawaoka.** 2001. Molecular basis for high virulence
580 of Hong Kong H5N1 influenza A viruses. *Science.* **293**:1840-1842. doi: 10.1126/science.1062882.
- 581 24. **Gabriel, G., B. Dauber, T. Wolff, O. Planz, H. D. Klenk, and J. Stech.** 2005. The viral

- polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc. Natl. Acad. Sci. U. S. A.* **102**:18590-18595. doi: 0507415102.
25. **Forni, D., G. Filippi, R. Cagliani, L. De Gioia, U. Pozzoli, N. Al-Daghri, M. Clerici, and M. Sironi.** 2015. The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses. *Sci. Rep.* **5**:14480. doi: 10.1038/srep14480.
26. **Loytynoja, A., and N. Goldman.** 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U. S. A.* **102**:10557-10562. doi: 0409137102.
27. **Penn, O., E. Privman, H. Ashkenazy, G. Landan, D. Graur, and T. Pupko.** 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* **38**:W23-8. doi: 10.1093/nar/gkq443.
28. **Privman, E., O. Penn, and T. Pupko.** 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* **29**:1-5. doi: 10.1093/molbev/msr177.
29. **Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost.** 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* **23**:1891-1901. doi: 10.1093/molbev/msl051.
30. **Sawyer, S.** 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**:526-538.
31. **Guindon, S., F. Delsuc, J. F. Dufayard, and O. Gascuel.** 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**:113-137. doi: 10.1007/978-1-59745-251-9_6; 10.1007/978-1-59745-251-9_6.
32. **Zhang, J., R. Nielsen, and Z. Yang.** 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**:2472-2479. doi: 10.1093/molbev/msi237.
33. **Yang, Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*

- 606 **24**:1586-1591. doi: 10.1093/molbev/msm088.
- 607 34. **Anisimova, M., and Z. Yang.** 2007. Multiple hypothesis testing to detect lineages under
608 positive selection that affects only a few sites. *Mol. Biol. Evol.* **24**:1219-1228. doi:
609 10.1093/molbev/msm042.
- 610 35. **Murrell, B., S. Weaver, M. D. Smith, J. O. Wertheim, S. Murrell, A. Aylward, K. Eren, T.**
611 **Pollner, D. P. Martin, D. M. Smith, K. Scheffler, and S. L. Kosakovsky Pond.** 2015. Gene-Wide
612 Identification of Episodic Selection. *Mol. Biol. Evol.* . doi: msv035.
- 613 36. **Pond, S. L., S. D. Frost, and S. V. Muse.** 2005. HyPhy: hypothesis testing using phylogenies.
614 *Bioinformatics.* **21**:676-679. doi: bti079.
- 615 37. **Anisimova, M., J. P. Bielawski, and Z. Yang.** 2002. Accuracy and power of bayes prediction
616 of amino acid sites under positive selection. *Mol. Biol. Evol.* **19**:950-958.
- 617 38. **Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, and S. L. Kosakovsky**
618 **Pond.** 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.*
619 **8**:e1002764. doi: 10.1371/journal.pgen.1002764.
- 620 39. **Delport, W., A. F. Poon, S. D. Frost, and S. L. Kosakovsky Pond.** 2010. Datamonkey 2010: a
621 suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics.* **26**:2455-2457. doi:
622 10.1093/bioinformatics/btq429.
- 623 40. **Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost.** 2007. An evolutionary-network model
624 reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput. Biol.* **3**:e231.
625 doi: 07-PLCB-RA-0289.
- 626 41. **Simonetti, F. L., E. Teppa, A. Chernomoretz, M. Nielsen, and C. Marino Buslje.** 2013.
627 MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.* **41**:W8-14. doi:
628 10.1093/nar/gkt427.
- 629 42. **Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer.** 2001. Predicting

- transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**:567-580. doi: 10.1006/jmbi.2000.4315.
43. **Yang, L., Z. Wu, X. Ren, F. Yang, J. Zhang, G. He, J. Dong, L. Sun, Y. Zhu, S. Zhang, and Q. Jin.** 2014. MERS-related betacoronavirus in *Vespertilio superans* bats, China. *Emerg. Infect. Dis.* **20**:1260-1262. doi: 10.3201/eid2007.140318.
44. **Tang, X. C., J. X. Zhang, S. Y. Zhang, P. Wang, X. H. Fan, L. F. Li, G. Li, B. Q. Dong, W. Liu, C. L. Cheung, K. M. Xu, W. J. Song, D. Vijaykrishna, L. L. Poon, J. S. Peiris, G. J. Smith, H. Chen, and Y. Guan.** 2006. Prevalence and genetic diversity of coronaviruses in bats from China. *J. Virol.* **80**:7481-7490. doi: 10.1128/JVI.01600-13.
45. **Corman, V. M., R. Kallies, H. Philipps, G. Gopner, M. A. Muller, I. Eckerle, S. Brunink, C. Drosten, and J. F. Drexler.** 2014. Characterization of a novel betacoronavirus related to middle East respiratory syndrome coronavirus in European hedgehogs. *J. Virol.* **88**:717-724. doi: 10.1128/JVI.01498-14.
46. **Corman, V. M., N. L. Ithete, L. R. Richards, M. C. Schoeman, W. Preiser, C. Drosten, and J. F. Drexler.** 2014. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. *J. Virol.* **88**:11297-11303. doi: 10.1128/JVI.01280-15.
47. **Wang, Y., D. Liu, W. Shi, R. Lu, W. Wang, Y. Zhao, Y. Deng, W. Zhou, H. Ren, J. Wu, Y. Wang, G. Wu, G. F. Gao, and W. Tan.** 2015. Origin and Possible Genetic Recombination of the Middle East Respiratory Syndrome Coronavirus from the First Imported Case in China: Phylogenetics and Coalescence Analysis. *Mbio.* **6**:10.1128/mBio.01280-15. doi: 10.1128/mBio.01280-15.
48. **Serrano, P., M. A. Johnson, M. S. Almeida, R. Horst, T. Herrmann, J. S. Joseph, B. W. Neuman, V. Subramanian, K. S. Saikatendu, M. J. Buchmeier, R. C. Stevens, P. Kuhn, and K.**

- 654 **Wuthrich.** 2007. Nuclear magnetic resonance structure of the N-terminal domain of nonstructural
655 protein 3 from the severe acute respiratory syndrome coronavirus. *J. Virol.* **81**:12049-12060. doi:
656 JVI.00969-07.
- 657 49. **Wu, A., Y. Wang, C. Zeng, X. Huang, S. Xu, C. Su, M. Wang, Y. Chen, and D. Guo.** 2015.
658 Prediction and biochemical analysis of putative cleavage sites of the 3C-like protease of Middle
659 East respiratory syndrome coronavirus. *Virus Res.* **208**:56-65. doi: 10.1016/j.virusres.2015.05.018.
- 660 50. **Hurst, K. R., C. A. Koetzner, and P. S. Masters.** 2013. Characterization of a critical
661 interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral
662 replicase-transcriptase complex. *J. Virol.* **87**:9159-9172. doi: 10.1128/JVI.01275-13.
- 663 51. **Xu, Y., L. Cong, C. Chen, L. Wei, Q. Zhao, X. Xu, Y. Ma, M. Bartlam, and Z. Rao.** 2009.
664 Crystal structures of two coronavirus ADP-ribose-1"-monophosphatases and their complexes with
665 ADP-Ribose: a systematic structural analysis of the viral ADRP domain. *J. Virol.* **83**:1083-1092.
666 doi: 10.1128/JVI.01862-08.
- 667 52. **Serrano, P., J. Ramajo, and E. Martinez-Salas.** 2009. Rescue of internal initiation of
668 translation by RNA complementation provides evidence for a distribution of functions between
669 individual IRES domains. *Virology.* **388**:221-229. doi: 10.1016/j.virol.2009.03.021.
- 670 53. **Oostra, M., M. C. Hagemeijer, M. van Gent, C. P. Bekker, E. G. te Lintelo, P. J. Rottier,**
671 **and C. A. de Haan.** 2008. Topology and membrane anchoring of the coronavirus replication
672 complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *J. Virol.*
673 **82**:12392-12405. doi: 10.1128/JVI.01219-08.
- 674 54. **Hagemeijer, M. C., I. Monastyrska, J. Griffith, P. van der Sluijs, J. Voortman, P. M. van**
675 **Bergen en Henegouwen, A. M. Vonk, P. J. Rottier, F. Reggiori, and C. A. de Haan.** 2014.
676 Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology.* **458**-
677 **459**:125-135. doi: 10.1016/j.virol.2014.04.027.

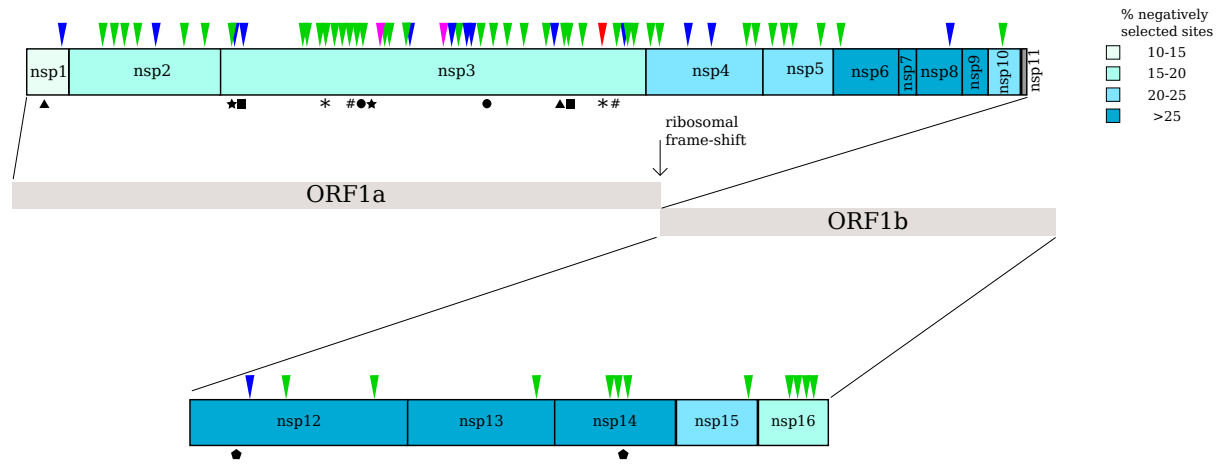
- 678 55. **Beachboard, D. C., J. M. Anderson-Daniels, and M. R. Denison.** 2015. Mutations across
679 murine hepatitis virus nsp4 alter virus fitness and membrane modifications. *J. Virol.* **89**:2080-2089.
680 doi: 10.1128/JVI.02776-14.
- 681 56. **Xu, X., Z. Lou, Y. Ma, X. Chen, Z. Yang, X. Tong, Q. Zhao, Y. Xu, H. Deng, M. Bartlam,**
682 **and Z. Rao.** 2009. Crystal structure of the C-terminal cytoplasmic domain of non-structural protein
683 4 from mouse hepatitis virus A59. *PLoS One.* **4**:e6217. doi: 10.1371/journal.pone.0006217.
- 684 57. **Tomar, S., M. L. Johnston, S. E. St John, H. L. Osswald, P. R. Nyalapatla, L. N. Paul, A. K.**
685 **Ghosh, M. R. Denison, and A. D. Mesecar.** 2015. Ligand-induced Dimerization of Middle East
686 Respiratory Syndrome (MERS) Coronavirus nsp5 Protease (3CLpro): IMPLICATIONS FOR nsp5
687 REGULATION AND THE DEVELOPMENT OF ANTIVIRALS. *J. Biol. Chem.* **290**:19403-19422.
688 doi: 10.1074/jbc.M115.651463.
- 689 58. **Stobart, C. C., A. S. Lee, X. Lu, and M. R. Denison.** 2012. Temperature-sensitive mutants and
690 revertants in the coronavirus nonstructural protein 5 protease (3CLpro) define residues involved in
691 long-distance communication and regulation of protease activity. *J. Virol.* **86**:4801-4810. doi:
692 10.1128/JVI.06754-11.
- 693 59. **Stobart, C. C., N. R. Sexton, H. Munjal, X. Lu, K. L. Molland, S. Tomar, A. D. Mesecar,**
694 **and M. R. Denison.** 2013. Chimeric exchange of coronavirus nsp5 proteases (3CLpro) identifies
695 common and divergent regulatory determinants of protease activity. *J. Virol.* **87**:12611-12618. doi:
696 10.1128/JVI.02050-13.
- 697 60. **Lugari, A., S. Betzi, E. Decroly, E. Bonnaud, A. Hermant, J. C. Guillemot, C. Debarnot, J.**
698 **P. Borg, M. Bouvet, B. Canard, X. Morelli, and P. Lecine.** 2010. Molecular mapping of the RNA
699 Cap 2'-O-methyltransferase activation interface between severe acute respiratory syndrome
700 coronavirus nsp10 and nsp16. *J. Biol. Chem.* **285**:33230-33241. doi: 10.1074/jbc.M110.120014.
- 701 61. **Bouvet, M., A. Lugari, C. C. Posthuma, J. C. Zevenhoven, S. Bernard, S. Betzi, I. Imbert,**

- 702 **B. Canard, J. C. Guillemot, P. Lecine, S. Pfefferle, C. Drosten, E. J. Snijder, E. Decroly, and X.**
703 **Morelli.** 2014. Coronavirus Nsp10, a critical co-factor for activation of multiple replicative
704 enzymes. *J. Biol. Chem.* **289**:25783-25796. doi: 10.1074/jbc.M114.577353.
- 705 62. **Donaldson, E. F., R. L. Graham, A. C. Sims, M. R. Denison, and R. S. Baric.** 2007. Analysis
706 of murine hepatitis virus strain A59 temperature-sensitive mutant TS-LA6 suggests that nsp10 plays
707 a critical role in polyprotein processing. *J. Virol.* **81**:7086-7098. doi: JVI.00049-07.
- 708 63. **Ma, Y., L. Wu, N. Shaw, Y. Gao, J. Wang, Y. Sun, Z. Lou, L. Yan, R. Zhang, and Z. Rao.**
709 2015. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc.*
710 *Natl. Acad. Sci. U. S. A.* **112**:9436-9441. doi: 10.1073/pnas.1508686112.
- 711 64. **Sawicki, S. G., D. L. Sawicki, D. Younker, Y. Meyer, V. Thiel, H. Stokes, and S. G. Siddell.**
712 2005. Functional and genetic analysis of coronavirus replicase-transcriptase proteins. *PLoS Pathog.*
713 **1**:e39. doi: 10.1371/journal.ppat.0010039.
- 714 65. **Decroly, E., C. Debarnot, F. Ferron, M. Bouvet, B. Coutard, I. Imbert, L. Gluais, N.**
715 **Papageorgiou, A. Sharff, G. Bricogne, M. Ortiz-Lombardia, J. Lescar, and B. Canard.** 2011.
716 Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-O-methyltransferase
717 nsp10/nsp16 complex. *PLoS Pathog.* **7**:e1002059. doi: 10.1371/journal.ppat.1002059.
- 718 66. **Lau, S. K., P. C. Woo, K. S. Li, Y. Huang, H. W. Tsoi, B. H. Wong, S. S. Wong, S. Y. Leung,**
719 **K. H. Chan, and K. Y. Yuen.** 2005. Severe acute respiratory syndrome coronavirus-like virus in
720 Chinese horseshoe bats. *Proc. Natl. Acad. Sci. U. S. A.* **102**:14040-14045. doi: 0506735102.
- 721 67. **Annan, A., H. J. Baldwin, V. M. Corman, S. M. Klose, M. Owusu, E. E. Nkrumah, E. K.**
722 **Badu, P. Anti, O. Agbenyega, B. Meyer, S. Oppong, Y. A. Sarkodie, E. K. Kalko, P. H. Lina, E.**
723 **V. Godlevska, C. Reusken, A. Seebens, F. Gloza-Rausch, P. Vallo, M. Tschapka, C. Drosten,**
724 **and J. F. Drexler.** 2013. Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and
725 Europe. *Emerg. Infect. Dis.* **19**:456-459. doi: 10.3201/eid1903.121503.

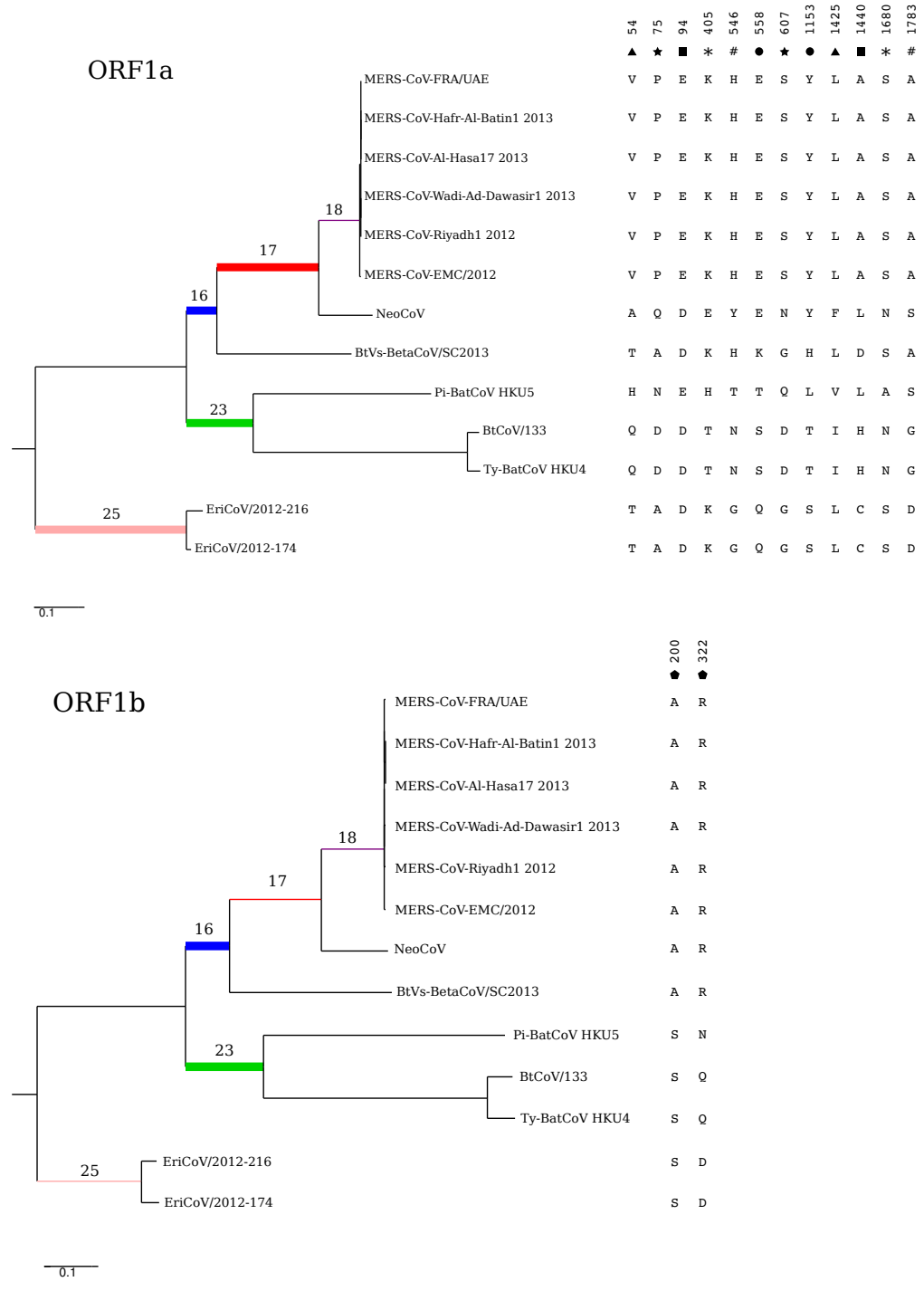
- 726 68. **Ithete, N. L., S. Stoffberg, V. M. Corman, V. M. Cottontail, L. R. Richards, M. C.**
727 **Schoeman, C. Drosten, J. F. Drexler, and W. Preiser.** 2013. Close relative of human Middle East
728 respiratory syndrome coronavirus in bat, South Africa. *Emerg. Infect. Dis.* **19**:1697-1699. doi:
729 10.3201/eid1910.130946.
- 730 69. **Anthony, S. J., R. Ojeda-Flores, O. Rico-Chavez, I. Navarrete-Macias, C. M. Zambrana-**
731 **Torrelío, M. K. Rostal, J. H. Epstein, T. Tipps, E. Liang, M. Sanchez-Leon, J. Sotomayor-**
732 **Bonilla, A. A. Aguirre, R. Avila-Flores, R. A. Medellin, T. Goldstein, G. Suzan, P. Daszak, and**
733 **W. I. Lipkin.** 2013. Coronaviruses in bats from Mexico. *J. Gen. Virol.* **94**:1028-1038. doi:
734 10.1099/vir.0.049759-0.
- 735 70. **Gharib, W. H., and M. Robinson-Rechavi.** 2013. The branch-site test of positive selection is
736 surprisingly robust but lacks power under synonymous substitution saturation and variation in GC.
737 *Mol. Biol. Evol.* **30**:1675-1686. doi: 10.1093/molbev/mst062.
- 738 71. **Bhatt, S., T. T. Lam, S. J. Lycett, A. J. Leigh Brown, T. A. Bowden, E. C. Holmes, Y. Guan,**
739 **J. L. Wood, I. H. Brown, P. Kellam, Combating Swine Influenza Consortium, and O. G.**
740 **Pybus.** 2013. The evolutionary dynamics of influenza A virus adaptation to mammalian hosts.
741 *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**:20120382. doi: 10.1098/rstb.2012.0382.
- 742 72. **Gong, L. I., and J. D. Bloom.** 2014. Epistatically interacting substitutions are enriched during
743 adaptive protein evolution. *PLoS Genet.* **10**:e1004328. doi: 10.1371/journal.pgen.1004328.
- 744 73. **Longdon, B., M. A. Brockhurst, C. A. Russell, J. J. Welch, and F. M. Jiggins.** 2014. The
745 evolution and genetics of virus host shifts. *PLoS Pathog.* **10**:e1004395. doi:
746 10.1371/journal.ppat.1004395.
- 747 74. **Bedhomme, S., J. Hillung, and S. F. Elena.** 2015. Emerging viruses: why they are not jacks of
748 all trades? *Curr. Opin. Virol.* **10**:1-6. doi: 10.1016/j.coviro.2014.10.006.
- 749 75. **Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. M. Whitcomb, C. J.**

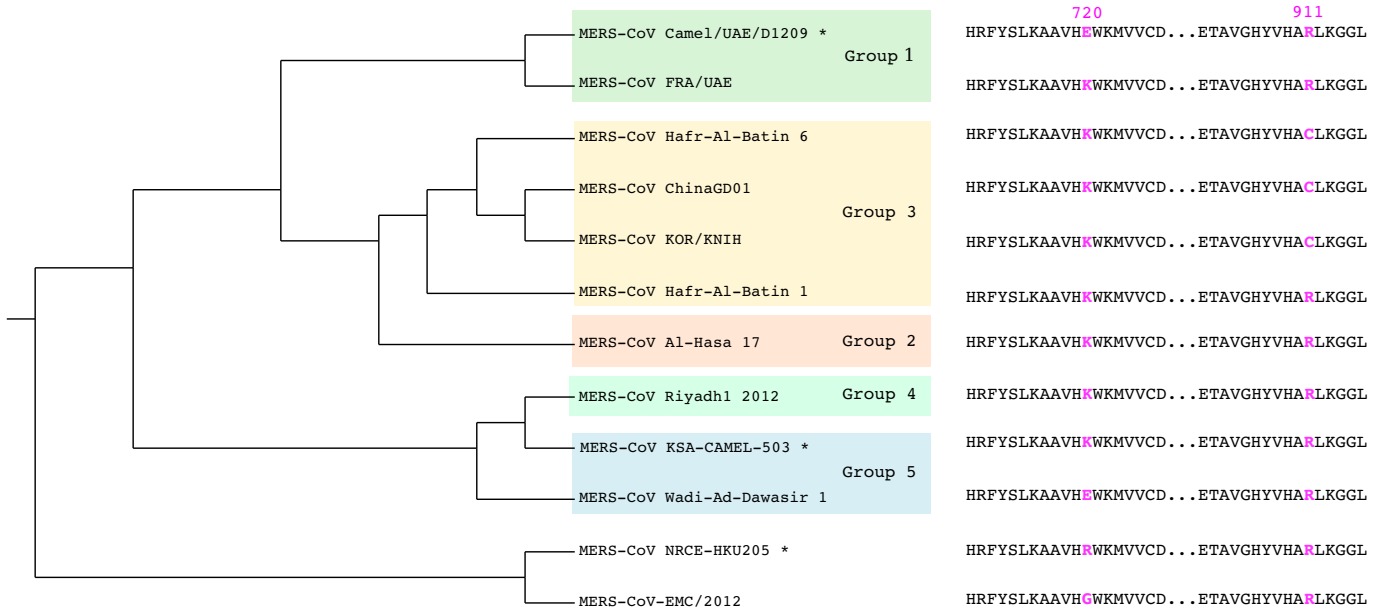
- 750 **Petropoulos, and S. Bonhoeffer.** 2011. A systems analysis of mutational effects in HIV-1 protease
751 and reverse transcriptase. *Nat. Genet.* **43**:487-489. doi: 10.1038/ng.795.
- 752 76. **Parera, M., and M. A. Martinez.** 2014. Strong epistatic interactions within a single protein.
753 *Mol. Biol. Evol.* **31**:1546-1553. doi: 10.1093/molbev/msu113.
- 754 77. **Neverov, A. D., S. Kryazhimskiy, J. B. Plotkin, and G. A. Bazykin.** 2015. Coordinated
755 Evolution of Influenza A Surface Proteins. *PLoS Genet.* **11**:e1005404. doi:
756 10.1371/journal.pgen.1005404.
- 757 78. **Kryazhimskiy, S., J. Dushoff, G. A. Bazykin, and J. B. Plotkin.** 2011. Prevalence of epistasis
758 in the evolution of influenza A surface proteins. *PLoS Genet.* **7**:e1001301. doi:
759 10.1371/journal.pgen.1001301.
- 760 79. **Tsetsarkin, K. A., R. Chen, G. Leal, N. Forrester, S. Higgs, J. Huang, and S. C. Weaver.**
761 2011. Chikungunya virus emergence is constrained in Asia by lineage-specific adaptive landscapes.
762 *Proc. Natl. Acad. Sci. U. S. A.* **108**:7872-7877. doi: 10.1073/pnas.1018344108.
- 763 80. **Stokes, H. L., S. Baliji, C. G. Hui, S. G. Sawicki, S. C. Baker, and S. G. Siddell.** 2010. A new
764 cistron in the murine hepatitis virus replicase gene. *J. Virol.* **84**:10148-10158. doi:
765 10.1128/JVI.00901-10.
- 766 81. **Sawyer, S. L., and N. C. Elde.** 2012. A cross-species view on viruses. *Curr. Opin. Virol.* **2**:561-
767 568. doi: 10.1016/j.coviro.2012.07.003.
- 768

A



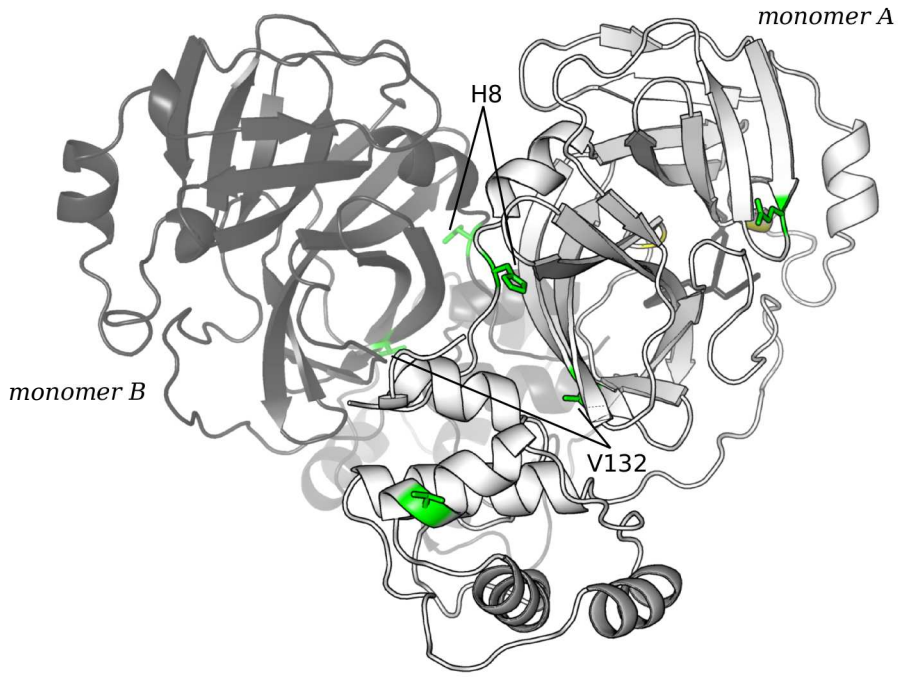
B





The diagram illustrates the Lumenal loop (ZF) of a protein, showing its structure and the probability distribution of residues. The protein is oriented with the Lumen at the top and the Cytosol at the bottom. The Lumenal loop (ZF) is highlighted in yellow and contains four transmembrane helices (TMH1, TMH2, TMH3, and TMH4). A1386 is marked with a blue arrowhead. A N-glycosylation site is indicated by a black arrowhead. The probability distribution of residues is shown as a line graph below the protein structure, with the Y-axis representing Probability (0.0 to 1.0) and the X-axis representing Residues (1235 to 1535). The distribution shows four distinct peaks corresponding to the transmembrane helices.

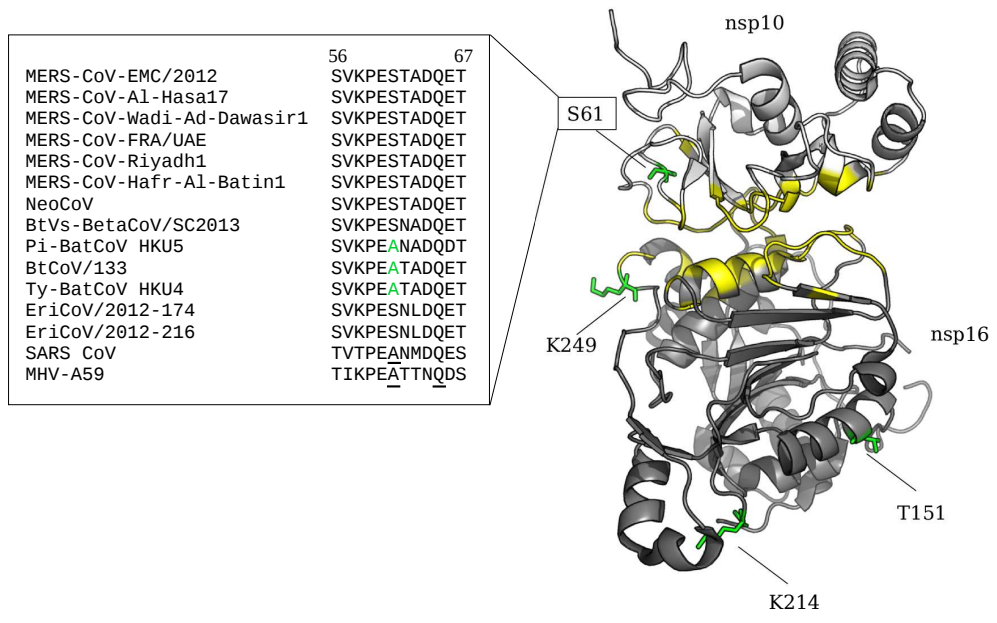
A



B

	1	32	121	149
MERS-CoV-EMC/2012	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
MERS-CoV-Al-Hasa17	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
MERS-CoV-Wadi-Ad-Dawasir1	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
MERS-CoV-FRA/UAE	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
MERS-CoV-Riyadh1	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
MERS-CoV-Hafr-Al-Batin1	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
NeoCoV	SGLVKMSHPSGDVEACMVQVTCG...	YNGRPTGFTTVMRPNYTIKGSFLCGSCG		
BtVs-BetaCoV/SC2013	SGLVKMSHPSGAVEACMVQVTCG...	YNGRPTGFTTVMRPNHTIKGSFLCGSCG		
Pi-BatCoV HKU5	SGLVKMAAPSGVVENCIVQVTCG...	YNGRPTGVFMVNMQRNSTIKGSFLCGSCG		
BtCoV/133	SGLVKMSAPSGAVENCIVQVTCG...	YNGKPTGVFTVNLRHNSTIKGSFLCGSCG		
Ty-BatCoV HKU4	SGLVKMSAPSGAVENCIVQVTCG...	YNGKPTGVFTVNLRHNSTIKGSFLCGSCG		
EriCoV/2012-174	SGLVKMAHPSGAVEQCIQVQVTCG...	YNGRPSGTYTVVMRPNSTIKGSFLCGSCG		
EriCoV/2012-216	SGLVKMAHPSGAVEQCIQVQVTCG...	YNGRPSGTYTVVMRPNSTIKGSFLCGSCG		
SARS CoV	SGFRKMAFSPSGKVEGCMVQVTCG...	YNGSPSGVYQCAMPNHTIKGSFLNGSCG		
MHV-A59	SGIVKMVSPTSKEPCIVSVTYG...	YNGRPQGAHFVILRSSHTIKGSFLCGSCG		

A



B

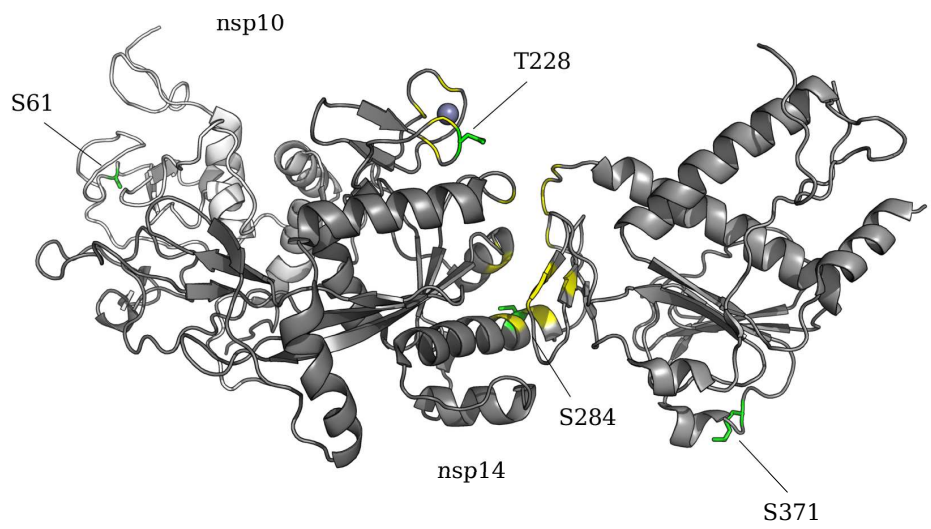


Table 1. Likelihood ratio test statistics for branch-site tests.

Region	Foreground branch ^a	-2ΔlnL ^b (MA vs MA1) ^c	MA vs MA1 <i>p</i> value (FDR corrected <i>p</i> value)	-2ΔlnL ^b (BUSTED)	BUSTED <i>p</i> value	N. of sites identified by BEB and BUSTED
ORF1a						
	Branch 18	4.35	0.0369 (0.0369)	0.50	0.778	-
	Branch 17	49.18	2.34x10 ⁻¹² (3.90x10 ⁻¹²)	23.18	9.25x10 ⁻⁶	1 (nsp3)
	Branch 16	75.22	4.21x10 ⁻¹⁸ (1.05x10 ⁻¹⁷)	85.9	2.58x10 ⁻¹⁹	13 (1 nsp1, 1 nsp2, 8 nsp3, 2 nsp4, 1 nsp8)
	Branch 23	93.39	4.29x10 ⁻²² (2.14x10 ⁻²¹)	71.76	2.61x10 ⁻¹⁶	41 (6 nsp2, 25 nsp3, 4 nsp4, 4 nsp5, 1 nsp6, 1 nsp10)
	Branch 25	26.84	2.20x10 ⁻⁷ (2.75x10 ⁻⁷)	7.52	0.023	-
ORF1b						
	Branch 18	0.78	0.38 (0.475)	-	-	-
	Branch 17	8.53	0.0035 (0.0058)	2.8	0.246	-
	Branch 16	12.70	3.6x10 ⁻⁴ (0.0009)	17.36	1.70x10 ⁻⁴	1 (nsp12)
	Branch 23	37.83	7.73x10 ⁻¹⁰ (3.86x10 ⁻⁹)	31.94	1.16x10 ⁻⁷	11 (2 nsp12, 1 nsp13, 3 nsp14, 1 nsp15, 4 nsp16)
	Branch 25	0.23	0.63 (0.63)	-	-	-

^a Branches are numbered as in figure 1.
^b 2ΔlnL is twice the difference of the natural logs of the maximum likelihood of the models being compared.
^c MA and MA1 are branch-site models that assume four classes of sites: the MA model allows a proportion of codons to have ω ≥ 1 on the foreground branches, whereas the MA1 model does not.

Table 2. Likelihood ratio test statistics for models of variable selective pressure among sites.

region	LRT model	$-2\Delta\ln L^c$	p value	% of sites (average ω)	Positively selected sites (BEB and MEME)
nsp3					
	M1a vs M2a ^a	15.25	0.00048	0.6 (17.24)	G720, R911
	M7 vs M8 ^b	14.15	0.00085	1.6 (11.59)	

^a M1a is a nearly neutral model that assumes one ω class between 0 and 1, and one class with $\omega=1$; M2a (positive selection model) is the same as M1a plus an extra class of $\omega >1$.

^b M7 is a null model that assumes that $0 < \omega < 1$ is beta distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with $\omega >1$.

^c $2\Delta\ln L$: twice the difference of the natural logs of the maximum likelihood of the models being compared.